基于自学习的汉语开放域命名实体边界识别

付瑞吉，秦　兵，刘　挺

（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘　要： 命名实体识别是自然语言处理领域的一个重要任务，为许多上层应用提供支持。本文主要研究汉语开放域命名实体边界的识别。由于目前该任务尚缺乏训练语料，而人工标注语料的代价又太大，本文首先基于双语平行语料和英语句法分析器自动标注了一个汉语专有名词语料，另外基于汉语依存树库生成了一个名词复合短语语料，然后使用自学习方法将这两部分语料融合形成命名实体边界识别语料，同时训练边界识别模型。实验结果表明自学习的方法可以提高边界识别的准确率和召回率。

关键词： 开放域命名实体识别；自学习；训练语料融合

中图分类号： TP391.12　　**文献标识码：** A　　**文章编号：** 2095－2163(2014)04－0001－05

Chinese Open－domain Named Entity Boundary Identification based on A Self－Training Method

FU Ruiji，QIN Bing，LIU Ting

（School of Computer Science and Technology，Harbin Institute of Technology，Harbin 150001，China）

Abstract: Named entity recognition is an important task in the domain of Natural Language Processing，which plays an important role in many applications. This paper focuses on the boundary identification of Chinese open－domain named entities. Because the shortage of training data and the huge cost of manual annotation，the paper proposes a self－training approach to identify the boundaries of Chinese open－domain named entities in context. Due to the lack of training data，the paper firstly generates a large scale Chinese proper noun corpus based on parallel corpora，and also transforms a Chinese dependency tree bank to a noun compound training corpus. Subsequently，the paper proposes a self－training－based approach to combine the two corpora and train a model to identify boundaries of named entities. The experiments show the proposed method can take full advantage of the two corpora and improve the performance of named entity boundary identification.

Key words: Open－domain Named Entity Recognition；Self－training；Training Corpus Combination

0　引　言

命名实体是文本中承载信息的重要语言单位，命名实体的识别和分类在信息抽取、开放域问答、信息检索以及机器翻译等领域都占有非常重要的地位。输入自然语言文本，命名实体识别的任务在于将其中事物的名称标记出来并给予适当的语义类别。传统命名实体由于类别有限，并不能满足自然语言处理领域上层任务的全部需求，因此本文专注于开放域命名实体边界的识别的研究。

传统命名实体识别的主流方法是统计机器学习方法，使用标注好的训练集训练模型，然后用训练好的模型来进行命名实体的识别，并且大多数采用序列标注的方法，可以一次性将边界和类别都标出。但对于开放域命名实体来说，由于涉及的领域非常多，类型多且无法预知，所以人工标注语料是不现实的。由于英语中专有名词首字母通常大写，所以英语中专有名词的识别相对容易，有的研究直接会将首字母大写的单词串作为命名实体候选[1]。因此，本文转而利用英语的短语结构句法分析，借助少量规则标注专有名词短语，再通过双语平行语料将边界信息映射到汉语端[2]，从而实现命名实体边界识别语料的自动标注。但由于开放域命名实体的范围更大，一些命名实体在英语中并没有被标为专有名词，例如“大规模杀伤性武器(weapon of mass destruction)”、“中国近代史(the modern history of China)”等。因此，研究中另外基于一个汉语依存树库，利用一些启发式的规则标注名词复合短语，随后即使用半指导的自学习方法将两部分语料融合并训练命名实体边界识别模型。

与传统自学习方法不同之处在于，传统的自学习方法是基于一个已标注的集合和一个未标注的集合进行的，而本文则是基于两个部分标注的语料。本文的方法大概分为以下几个步骤。首先，使用专有名词语料训练模型，自动识别名词复合短语语料中的专有名词；然后，将语料中原有的名词复合短语和自动标注的专有名词及短语融合，得到初始的命名实体边界训练语料；接着，即在初始训练语料上训练命名实体边界识别模型，识别专有名词短语语料中的命名实体，选择高置信度的实例加入到训练语料中，如此迭代直到模型

收稿日期：2014－05－09
基金项目：国家自然科学基金(61133012，61273321)；国家高技术研究发展计划(863)前沿技术研究项目(2012AA011102)。
作者简介：付瑞吉(1984－)，男，陕西府谷人，博士研究生，主要研究方向：自然语言处理、文本挖掘；
秦　兵(1968－)，女，陕西华阴人，博士，教授，博士生导师，主要研究方向：自然语言处理、文本挖掘、情感分析等；
刘　挺(1972－)，男，黑龙江哈尔滨人，博士，教授，博士生导师，主要研究方向：自然语言处理、文本挖掘、社会计算等。

的性能稳定为止。

综上所述，本文提出了一种基于自学习的语料融合及模型训练的方法，用于汉语开放域命名实体识别边界的识别。实验证明本文的自学习方法是有效的，在测试集上获得了最好的 $F1$ 值。

1 自学习方法介绍

自学习(self teaching)，或叫自训练(self training)是常用的半指导机器学习方法。学术界对自学习有两种主要的定义。第一种定义是“单一视角的弱指导算法”，由 Ng 和 Cardie(2003)提出[3]。按照这种定义，可使用 bagging 方法从训练数据中随机采样训练多个分类器，预测时使用投票的方法决定最终的类别。利用这些分类器预测未标注数据，将“全票通过”的数据加入到训练集中，重新训练一组分类器，如此迭代，直到分类性能稳定。各分类器均采用相同的视角(view，可以理解为特征)训练。第二种定义是“基于分类器自己的标注结果重新训练分类器的方法”，由 Clark 等人(2003)提出[4]。首先在一个小规模的已标注数据上训练模型，然后使用该模型自动处理未标注数据，选择置信度最高的一部分自动处理的数据加入到训练集合中。接着重新训练模型，如此迭代，直到模型性能不再发生改进为止。这种方法中，模型利用自己的预测结果指导自己训练，所以叫做“自学习”。高置信度的数据通常基于一个阈值来判断，高于这个阈值才可选作训练数据，如此做法的目的即在于避免错误被加强。本文中，采取了第二种定义。迄今为止，自学习方法已经成功应用于自然语言的多个处理任务中，包括传统命名实体识别[5]、词义消歧[6]、句法分析[7]等。

2 基于自学习方法的命名实体边界识别

经过分析发现，开放域命名实体大概包括专有名词和名词复合短语两部分。专有名词通常指事物特定的名词，如“姚明”、“中国”、“伊拉克战争”等。而名词复合短语则用来表示语义更加宽泛的事物的名称，如“大规模杀伤性武器”、“中国近代史”等，这些也属于开放域命名实体的范畴。当然，还有一些专有名词同时也是名词复合短语。因此，研究通过分别构建这两部分语料，再通过自学习方法融合语料并训练命名实体边界识别的模型。

2.1 训练语料构建

2.1.1 基于双语平行语料的汉语专有名词识别语料构建

在此，即基于中英双语平行语料，并通过上节提出的方法来构建汉语命名实体边界识别的语料库。在英语上，借用了短语结构句法分析工具来识别英语的专有名词短语，由于英语具有大小写特征，因此对专有名词的识别尤其具有先天的优势。图1给出了一个短语结构句法分析的例子，其中“Ming”和“Yao”都被标为了专有名词(NNP)，并且两者结合形成了一个更大的名词短语(NP)，“Houston”和“Rockets”也是类似的情况，只不过在构成更大名词短语的时候加入了定冠词 the。

本文设计了一些规则来标注专有名词短语，如表1所示，其中的 NNP 均可替换为 NNPS(复数形式的专有名词)。然后，仍使用上一节提出的方法将英语端的标记映射到汉语端，实现语料的标注。

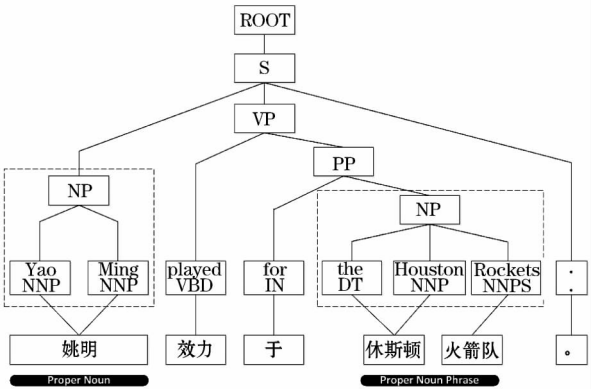

图1 基于双语平行语料和英语短语结构句法标注汉语专有名词语料示例

Fig. 1 An example of Chinese proper noun training data labeling based on an English – Chinese parallel corpus and an English parser

表1 基于短语结构句法分析结果标注专有名词的规则

Tab. 1 The rules of proper noun labeling based on parsing results

句法结构模式	标注结果
(NP (NNP w_1) ... (NNP w_n))	[w_1 ... w_n] NNP
(NP (DT the) (NNP w_1) ... (NNP w_n))	[the w_1 ... w_n] NNP
(NP (NNP w_1) ... (NNP w_{n-1}) (POS 's) (NN w_n))	[w_1 ... w_{n-1} 's w_n] NNP
(NP (NNP w_1) ... (NNP w_i)) (PP (IN of)) (NP (NNP w_{i+1}) ... (NNP w_n))	[w_1 .. w_i of w_{i+1} ... w_n] NNP
(NP (NNP w_1) ... (NNP w_i)) (CC and) (NNP w_{i+1}) ... (NNP w_n))	[w_1 ... w_i] NNP and [w_{i+1} ... w_n] NNP

2.1.2 基于依存树库的名词复合短语识别语料构建

名词复合短语是广泛存在于各种语言中的一种名词短语类型，从字面上可理解为两个或两个以上名词构成的名词性短语。名词复合短语在自然语言中出现频繁，广泛存在于各种文体中，其功能具有代表性，整体上相当于一个名词。由于语言的灵活性，不同语言中的名词复合短语有着不同的表述形式，甚至不同人对于同种语言相同形式的理解也不尽相同。本文从汉语语言特性和实际应用出发，结合前人研究工作，给出了涵盖范围更广的定义。

汉语名词复合短语是由体词及谓词等成分按顺序构成的单词序列，在语义上代表某一特定的实体或概念，短语的内部结构稳定。本文借鉴赵军和黄昌宁(1999)对汉语基本

名词短语的定义[8]，给出了汉语名词复合短语的形式化定义如下:

名词复合短语 = 限定语 + 核心词

核心词 → 名词|符号串|动词

限定语 → 名词|简称|符号串|动词|数词|量词|形容词

按照这个定义，“中国足球联赛”、“北京 101 中学”、“自然语言处理”、“第四次中东战争”等都属于名词复合短语。汉语属于意合语言，在名词短语的构成上非常灵活，一些动词也可以作为名词复合短语的组成成分，比如“中国驻俄罗斯大使馆”中的“驻”、“电影发行公司”中的“发行”、“未成年人保护法”中的“保护”等。而且，汉语不像英语有丰富的词形变换，同一个词义会有名词和动词的不同形式，因此增加了名词复合短语的识别难度。核心词一般出现在短语的末尾位置，如“哈尔滨工业大学”，“大学”为该短语的核心词。限定词可以由多个词组成，而核心词则一般为单个词。

根据这些特点，文中基于一些规则将汉语依存树库中的名词复合短语标出。汉语依存树库是人工标注了句子内词语之间依存关系的语料库，包括修饰关系、主谓关系、动宾关系等。规则的主要思想如下:

名词复合短语为连续串，两个短语不能重叠或嵌套;

通常两个名词复合短语不直接相连;

大部分核心词为名词，也有少量符号串和动词，形容词、副词、助词等其他词性的词语不能作核心词;

通常名词复合短语内部词语由修饰依存关系相连。

如图 2 展示了一个例子，在此即可依照上述规则标注名词复合短语“法国大革命战争”。

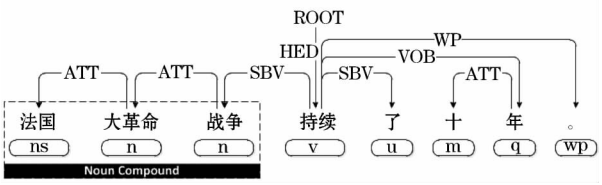

图 2 基于汉语依存树库标注名词复合短语语料示例

Fig. 2 An example of noun compound training data labeling based on a Chinese dependency treebank

2.2 基于自学习的命名实体边界识别模型训练

由于上述生成的两个语料是各有侧重点的，一个侧重于专有名词的标注，另一个侧重于名词复合短语的标注。在本小节，将采用自学习的方法融合这两部分语料，并训练命名实体边界识别的模型，方法框架如图 3 所示。

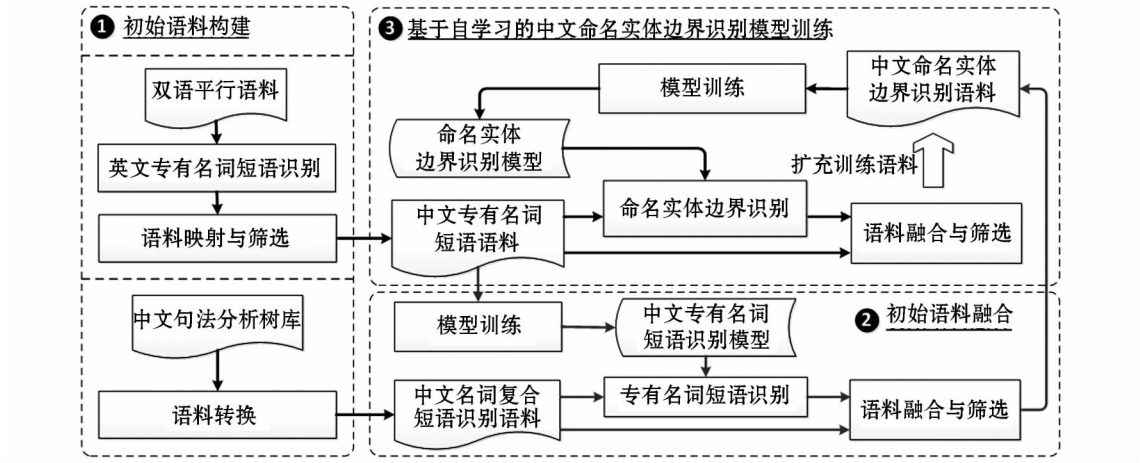

图 3 基于自学习的命名实体边界识别模型训练框架图

Fig. 3 The framework of model training of NE boundaries identification based on self – training

由图 3 可知，整个方法分为初始语料构建、初始语料融合和基于自学习的模型训练三个步骤。具体过程论述如下。

(1) 初始语料的构建: 这一步是基于双语平行语料和汉语句法树库分别构建专有名词短语语料和名词复合短语语料，更多细节可详见上一节。

(2) 初始语料融合: 利用专有名词短语语料训练序列标注模型，本文采用条件随机域模型(CRF)，并利用该模型对名词复合短语语料进行自动标注。标注后，即选取高质量的标注结果和原有的名词复合短语语料进行融合，得到开放域命名实体边界识别的语料。融合时如果遇到嵌套情况，则保留较长的命名实体; 如果遇到重叠的情况，则丢弃当前的句子，保证语料的质量。

(3) 基于自学习的模型训练: 在获得了一个小规模的命名实体边界识别的语料(称为初始语料)后，再通过自学习的方法逐步将专有名词语料融合进来，形成一个更大规模的语料。其中的自学习是一个迭代增强的过程: 首先利用初始语料训练命名实体边界识别模型，然后使用该模型标注专有名词短语语料; 接着就要选取高置信度的实例作为初始语料的补充，扩充后的语料又可以用来训练新的模型，如此迭代直到模型性能稳定为止。

3 实验结果与分析

3.1 实验数据

在语料构建方面，本文选取双语平行语料 LDC2003E14 和斯坦福的短语结构句法分析工具来生成汉语专有名词短语训练语料，基于上一节中的方法，最终获得 145 747 句专有名词短语训练语料。名词复合短语的生成则是基于哈尔滨工业大学社会计算与信息检索研究中心人工标注的 6 万句

汉语依存关系树库(HIT-IR Dependency Treebank)[9]。

在模型特征方面,本文则在上述的依存关系树库上统计动词依存关系的分值,并选用约400万百度百科词条及其开放类别信息挖掘命名实体的构成模式。

为了评测,进一步从OntoNotes 4.0语料中随机选取了8 789句标注,共包含19 315个命名实体,平均每个实体包含2.02个词。而且,由其中随机筛选1/5作为开发集,剩下的4/5为测试集。

3.2 自学习置信度阈值的选取

自学习的一个主要参数是每轮迭代增加新的训练语料时使用的置信度阈值θ。本文以句子为单位进行考察,如果命名实体边界识别模型对整句s序列标注的置信度$P_s > \theta$,则将s加入训练语料中,并从专有名词语料库中移除,否则将继续留存在专有名词语料库中,等待下一轮迭代时考察。在实验中,也尝试了改变θ的取值,观察自学习方法的学习曲线,图4显示了三个例子,即当θ的取值分别为0.65、0.92和0.95时的情况。

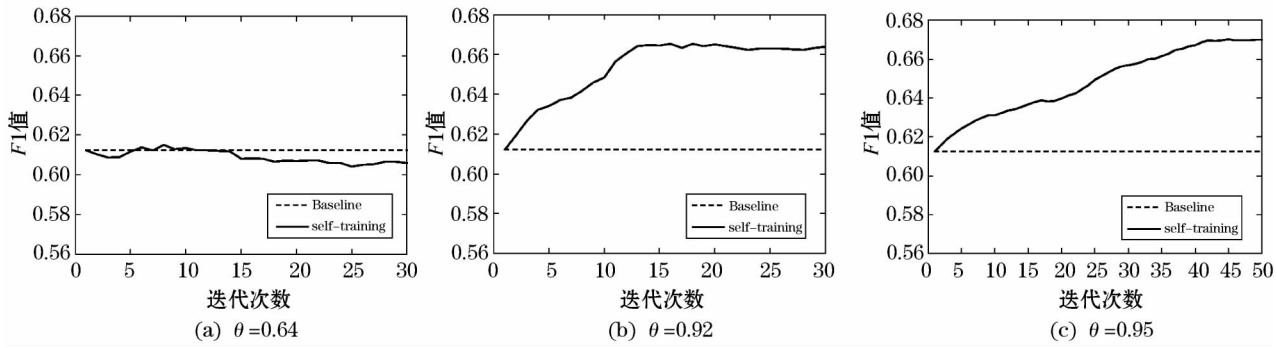

图4 不同置信度阈值下自学习方法在开发集上的学习曲线

Fig. 4 The learning curves of the self-training method under different confidence thresholds on the development dataset

由图4的三个学习曲线中,可以看到当θ取值较小时,自学习的方法并不能改进模型,性能反而有微弱的下降。这是因为阈值过小,使得新加入的语料中噪声过多,影响了模型的训练,而且不佳的模型会导致产生更多的噪声,形成恶性循环。而当θ取值过大时,自学习的收敛速度就会变慢,如当$\theta=0.95$时,就需要40次左右迭代,模型才能收敛;只有当$\theta=0.92$时,模型在13次迭代后即可收敛,并且最终的性能差距很小。因此,通过对训练速度和模型性能的综合评定,选取0.92为自学习置信度阈值。

其后,又在测试集上对本文涉及到的几个模型进行了对比,结果如表2所示。单纯使用专有名词语料(M_{NNP})或名词复合短语语料(M_{NC}),召回率都较低,需要结合两部分语料才能更加全面地识别命名实体。M_{NNP+NC}表示分别训练专有名词识别模型和名词复合短语识别模型,并且在识别结果上融合,融合的方法与本文初始语料融合的方法相同。M_{init}是基于融合后的命名实体初始语料训练模型的基线系统。M_{self}表示传统的自学习方法,即同样以初始命名实体识别语料启动,但使用完全未标注的生语料作为新的训练数据的来源。M_{self+}是本文的方法,可以看出文中的方法获得了最好的结果,比M_{self}提高了1.64%(显著性检验$p<0.01$),这就说明了来自于双语语料的专有名词短语信息的有效性。

表2 各命名实体边界识别模型之间的对比

Tab. 2 The comparison among NE boundary recognition models

模型	准确率	召回率	$F1$值
M_{NNP}	0.706 8	0.199 9	0.311 6
M_{NC}	0.555 4	0.498 1	0.525 2
M_{NNP+NC}	0.570 6	0.668	0.615 5
M_{init}	0.571 5	0.667 8	0.615 9
M_{self}	0.636 1	0.653 9	0.644 9
M_{self+}	0.641 3	0.682 6	0.661 3

4 结束语

本文针对开放域命名实体边界识别问题,提出了基于自学习的语料融合和模型训练方法。首先分别基于双语平行语料和汉语依存树库自动标注汉语专有名词语料和名词复合短语语料。然后基于自学习的方法将这两部分语料互补融合,形成命名实体边界识别的语料,同时在此基础上训练边界识别模型。实验证明了自学习方法的有效性,在开放域的测试语料上,本方法得到了最好的$F1$值0.661 3。

参考文献:

[1] EVANS R. A framework for named entity recognition in the open domain [J]. Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003, 2004, 260: 267 - 274.

[2] FU Ruiji, QIN Bing, LIU Ting. Exploiting multiple sources for open-domain hypernym discovery [C] // Proceedings of EMNLP 2013, 2013: 1224 - 1234.

[3] NG V, CARDIE C. Weakly supervised natural language learning without redundant views [C] //Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, 1: 94 - 101.

[4] CLARK S, CURRAN J R, OSBORNE M. Bootstrapping POS taggers using unlabelled data [C] // Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, 2003, 4: 49 - 55.

[5] KOZAREVA Z, BONEV B, MONTOYO A. Self-training and co-training applied to Spanish named entity recognition [M]. MICAI 2005: Advances in Artificial Intelligence. Springer, 2005: 770 - 779.

(下转第8页)

存在,则需要增加检测门限,这就将导致检测算法效率的降低。

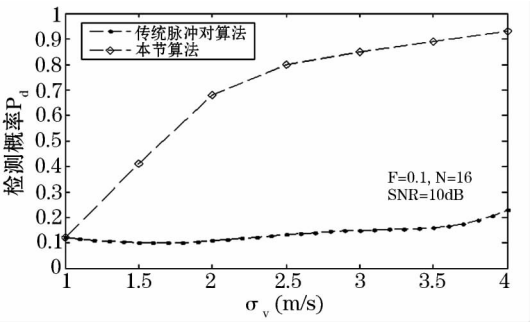

图7 $SNR=10$dB 下的检测性能

Fig. 7 Performation detection of $SNR=10$dB

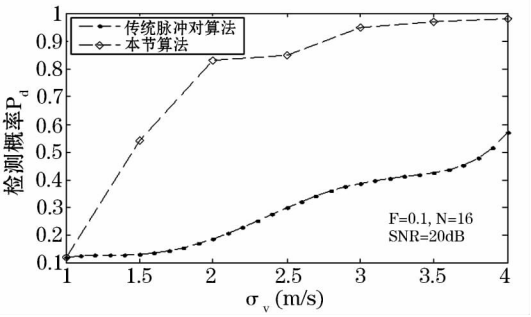

图8 $SNR=20$dB 下的检测性能

Fig. 8 Performation detection of $SNR=20$dB

图7是在 $SNR=10$dB($F=0.1$、$N=16$)下的检测概率特性图,其检测性能要优于传统的脉冲对检测算法,具体上升了48.92%;图8是在 $SNR=20$dB($F=0.1$、$N=16$)下的检测概率特性图,其检测性能要优于传统的脉冲对检测算法,实际上升了55.34%。同时,通过比较在不同信噪比下的检测概率,发现 $SNR=20$dB 下的检测性能要优于 $SNR=10$dB 下的性能,即提升了9.24%。

4 结束语

本文在分析传统的脉冲对湍流检测方法的基础上,提出了一种新的湍流检测算法。并经实验验证,提出的新算法的检测概率要优于传统的检测方法。在低 SNR 下,本文提出的湍流检测算法的性能将更为优秀;这是因为传统的脉冲对湍流检测算法主要是检测相关因子的减少和非相干噪声,这就意味着为了得到相同的虚警率,其检测门限将要设置一定的增加,同时也进一步说明了脉冲对算法在 $SNR=10$dB 或更小的情形下其检测效率会急剧减小的原因所在。

参考文献:

[1] MASON M S, WOOD G S, FLETCHER D F. Numerical simulation of downburst winds [J]. Journal of Wind Engineering and Industrial Aerodynamics, 2009, 97(11): 523 – 539.

[2] 刘小洋,李勇,程宇峰. 机载脉冲多普勒雷达湍流信号的仿真分析[J]. 系统工程与电子技术, 2012, 34(5): 920 – 924.

[3] MAZURA I V, YANOVSKY F J. Modeling of relationship between differential doppler velocity and turbulence [J]. Telecommunication and Radio Engineering, 2007, 66(12): 1113 – 1121.

[4] LIGTHART L P, YANOVSKY F J, PROKOPENKO I G. Adaptive algorithms for radar detection of turbulent zones in clouds and precipitation [J]. IEEE Transactions on aerospace and electronic systems, 2003, 39(1): 357 – 367.

[5] SANDALIDIS H G, TSIFTSIS T A, KARAGIANNIDIS G K, et al. BER performance of FSO links over strong atmospheric turbulence channels with pointing errors [J]. IEEE Communications Letters, 2008, 12(1): 44 – 46.

[6] 李勇,刘小洋,程宇峰. 机载雷达三维空间湍流场产生与仿真分析[J]. 系统工程与电子技术, 2013, 35(6): 1193 – 1198.

[7] NICOLA D D. Steady homogeneous turbulence in the presence of an average velocity gradient [J]. International Journal of Engineering Science, 2012, 51: 74 – 89.

[8] HUI M C H, LARSEN A, XIANG H F. Wind turbulence characteristics study at the Stonecutters Bridge site: Part II: Wind power spectra, integral length scales and coherences [J]. Journal of Wind Engineering and Industrial Aerodynamics, 2009, 97: 48 – 59.

[9] DANAILA L, ANTONIA R A, BURATTINI P. Comparison between kinetic energy and passive scalar energy transfer in locally homogeneous isotropic turbulence [J]. Nonlinear Phenomena: Physica D, 2012, 241: 224 – 231.

[10] THOMAS R, CHRISTIAN B, PIERRE M, et al. Generation of correlated stress time histories from continuous turbulence Power Spectral Density for fatigue analysis of aircraft structures [J]. International Journal of Fatigue, 2012, 42: 147 – 152.

[11] ZHU Xiang, MILANFAR, PEYMAN. Removing atmospheric turbulence via space – invariant deconvolution [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 157 – 170.

[12] OMAR O, LI Xin, MUBARAK S. Simultaneous Video stabilization and moving object detection in turbulence [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(2): 450 – 462.

(上接第4页)

[6] MIHALCEA R. Co – training and self – training for word sense disambiguation [C] //Proceedings of the Conference on Computational Natural Language Learning (CoNLL – 2004), 2004.

[7] McClosky D, Charniak E, Johnson M. Effective self – training for parsing [C]. Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics. 2006: 152 – 159.

[8] 赵军, 黄昌宁. 汉语基本名词短语结构分析模型 [J]. 计算机学报, 1999, 22(2): 141 – 146.

[9] LIU Ting, MA Jinshan, LI Sheng. Building a dependency treebank for improving Chinese parser [J]. Journal of Chinese Language and Computing, 2006, 16(4): 207 – 224.